AD_____

GRANT NO: DAMD17-94-J-4456

TITLE: Computer-Aided Mammography Using Automated Feature
Extraction for the Detection and Diagnosis of Breast Cancer

PRINCIPAL INVESTIGATOR(S):   Joseph Y. Lo

CONTRACTING ORGANIZATION: Duke University Medical Center
                             Durham, North Carolina   27710

REPORT DATE: October 12, 1995

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
                   Fort Detrick, Maryland   21702-5012

DISTRIBUTION STATEMENT: Approved for public release;
                          distribution unlimited

The views, opinions and/or findings contained in this report are
those of the author(s) and should not be construed as an official
Department of the Army position, policy or decision unless so
designated by other documentation.

REPORT DOCUMENTATION PAGE

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE 12 Oct 1995	3. REPORT TYPE AND D Annual Report (15 Sep 94 - 14 Sep 95)

4. TITLE AND SUBTITLE
Computer-Aided Mammography Using Automated Feature Extraction for the Detection and Diagnosis of Breast Cancer

5. FUNDING NUMBERS
DAMD17-94-J-4456

6. AUTHOR(S)

Joseph Y. Lo

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Duke University Medical Center
Durham, North Carolina 27710

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, MD 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

We developed artificial neural network (ANN) techniques to predict breast lesion malignancy based on mammographic features extracted by radiologists. The 3-layer backpropagation ANNs were trained and tested using the round robin technique and evaluated by ROC (receiver operating characteristic) analysis. Using all 11 available features from 206 patients, the ANN performed with ROC area A_z of 0.84 ± 0.03, which was not significantly different from the expert radiologists' A_z of 0.85 ± 0.03 (2-tailed p-value = 0.54). We then ranked the importance of individual features to reduce the number of ANN input features. The resulting 6-feature ANN had A_z of 0.86 ± 0.03 which was still not significantly different than that of the expert radiologists with $p = 0.34$. The result was an optimally simplified ANN for merging features to predict breast lesion malignancy. In the following years, work will focus on automated extraction of those features to feed into the ANN inputs, thus producing a fully automated computer-aided diagnosis system.

14. SUBJECT TERMS

computer-aided diagnosis, image processing, breast cancer, mammography

15. NUMBER OF PAGES
21

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

( ) Where copyrighted material is quoted, permission has been obtained to use such material.

( ) Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

(JH) Citations of commercial organizations and trade names in this report do not constitute an official Department of the Army endorsement or approval of the products or services of these organizations.

( ) In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Animal Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

( ) For the protection of human subjects, the investigator(s) have adhered to policies of applicable Federal Law 32 CFR 219 and 45 CFR 46.

( ) In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

_____
Principal Investigator's Signature

10/12/95
_____
Date

Accesion For		
NTIS CRA&I		☒
DTIC TAB		☒
Unannounced		☐
Justification		
By		
Distribution /		
Availability Codes		
Dist	Avail and / or Special	
A-1		

Table of Contents

1.   Introduction

1.1. Significance of diagnostic problem

In the U.S. in 1994, there were approximately 182,000 new cases and 46,000 deaths due to breast cancer, making it second only to lung cancer as the cause of cancer death among women [1]. Mammography is the modality of choice for early detection of breast cancer and can significantly decrease the mortality for women undergoing screening [2,3]. Evaluating mammograms remains a challenging task to radiologists, however, as they consider many radiographic and non-radiographic features in order to decide whether a lesion is benign or whether it should be followed or biopsied. Although mammography is very sensitive, there are a large number of false-positive biopsies. Of women with radiographically-suspicious, nonpalpable lesions who are sent to biopsy, only 15 to 34% actually have a malignancy by histologic diagnosis [4,5].

1.2. Potential of the proposed technique

This study seeks to improve the diagnosis and treatment of breast cancer by reducing the cost and morbidity of unnecessary biopsies. Cost is a major obstacle to widespread acceptance of mammographic screening [6]. It has been shown that surgeon's fees and biopsy costs account for over half the cost of detecting small breast cancers in a screening population [7]. Preventing unnecessary biopsies is therefore one of the most important ways to improve the efficacy of mammographic screening. Many previous reports have discussed the need to reduce the number of benign biopsies [8,9].

To improve early diagnosis, we propose an automated computer-aided diagnosis (CADx) system for mammography. The system will perform automated feature extraction from mammograms using artificial neural network (ANN) and other image processing techniques, then predict the outcome of biopsy (benign vs. malignant). The intent is to identify probably benign lesions for which biopsies may be spared. This study will potentially provide an accurate, consistent aide for the early diagnosis of breast cancer.

A successful mammography CADx system consists of two stages: (1) automated extraction of various features from the mammogram, potentially by different methods appropriate for each task, and (2) accurate "merging" of those features by computer algorithms to produce the diagnosis. Automated feature extraction can improve the accuracy, specificity, consistency, efficiency, and accessibility of breast cancer diagnosis.

1.3. Computer-aided diagnosis using artificial neural networks

In medical imaging, CADx systems provide radiologists with information from computerized analysis of images or image features, thus helping radiologists detect or diagnose diseases more accurately, easily, and consistently [10,11]. CADx has been applied to such varied problems as interstitial disease [12,13], cardiomegaly [14], pneumothorax [15], lung nodules [16-18], nuclear medicine lesions [19], and pulmonary embolism [20]. In mammography specifically, there have been numerous reports on computerized detection [21-27] or diagnosis [28-33] of breast cancer. Although both are generally considered CADx systems, detection systems locate suspicious lesions in an image, while diagnosis systems such as this study determine whether those lesions are benign or malignant.

This study focuses on the use of artificial neural networks (ANNs) which are computer models inspired by the structure and function of biological neural networks, such as the cerebral cortex of the human brain. Most ANNs are characterized by multiple, simple computing elements or *neurons* that work in parallel. The neurons interact globally through connections that have strengths or *weights*, and together they can duplicate aspects of human intelligence while incorporating the processing power of computers [34]. The classification rules are not defined *a priori*. instead the network is trained by presenting it with medical findings and final diagnoses from many patients. The network "learns" by adapting its weights to improve its diagnosis for each patient, just as physicians become more experienced with time. Once trained, the network can generalize to new patients it has not seen before.

ANNs are very useful in handling complex decision tasks such as those involved in medical diagnoses, where multiple findings are subtly related in ways which are often difficult to express in the form of diagnostic criteria. The networks can capture such relationships between the input findings to generate robust outputs. ANNs solve problems empirically without requiring any prior knowledge of distribution functions or any type of statistical modeling, yet ANNs are able to duplicate solutions of statistical methods [35]. Finally, ANNs are always consistent, for they are not prone to human fatigue or bias.

1.4. Technical Objectives

The technical objectives pertaining to the *first* budget period are aims 1a, 1b, and 2a from the list of aims for the *entire* budget period shown below:
(1) *Identify an optimal subset of features that would provide adequate diagnostic performance.*
   1a. Retrain the features-to-diagnosis ANN using sub-groups of features. The goal is to maintain the sensitivity of the original network while keeping specificity reasonably high.
   1b. Encode the multiple-value features into binary "sub-features", then repeat step 1a to reduce the number of sub-features. The sub-features will be easier to extract by automated schemes.

(2) *Investigate conventional and ANN methods for extracting the optimal subset of features directly from mammograms.*
    2a. Implement established techniques which have demonstrated promise for extracting features belonging to our reduced feature set.
    2b. Investigate several ANN techniques for feature extraction, focusing on features which may be difficult to classify by conventional techniques in step 2a. For both 2a and 2b, evaluate these techniques by comparing the extracted features against radiologists' findings.

(3) *Evaluate the automated CAD system clinically.*
    3a. Implement the CAD system by feeding the best feature extraction techniques from step 2 into the best features-to-diagnosis ANN from step 1, and compare the resulting diagnosis against the biopsy result.
    3b. Evaluate the accuracy of the CAD system retrospectively by using patient records from our computerized mammography database.
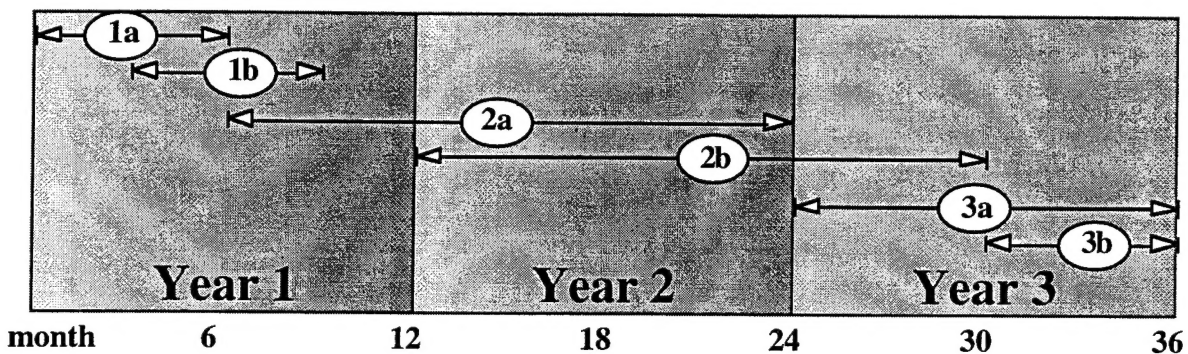

Figure 1. Time line for proposal project period.

In the following sections, we will report in detail on the progress in aim 1a. As will be explained, aim 1b will not be pursued. The preliminary results of aim 2 will be presented at a national conference during the second budget period [36], and discussion of those results will be reserved for the second annual report.

2.   Body

In the proposal, we reported some preliminary results from an artificial neural network (ANN) which predicted breast lesion malignancy based on mammographic findings extracted by radiologists [37]. The mammographic findings in that study were encoded using a lexicon which was at that time standard at our institution. Since then our institute has adopted a nationally-standardized lexicon. We investigated the use of this new lexicon to take advantage of its potential for general applicability of the CADx system. This work was presented at a national conference [38] and subsequently published in two parts in a peer-reviewed journal [39,40]. The ANN was developed using 206 patients who underwent excisional biopsy and pathologic diagnosis. The ANN was evaluated by receiver operating characteristic (ROC) analysis and its performance was compared to that of expert mammographers. That study was then extended by identifying an optimal subset of input features to simplify the network. This work was presented at two national conferences [41,42] and subsequently published in a peer-reviewed journal [43].

In the following sections, we will describe these studies in detail in the following order: the data collection and encoding scheme (2.1), the ANN architecture (2.2), the backpropagation algorithm used to train the ANN (2.3), the optimized reduction in the number of features (2.4), and the remaining technical objectives (2.5).

2.1. Data preparation

To provide the examples for supervised training of the neural network, the mammograms of 206 women with nonpalpable lesions were randomly selected for prospective evaluation from studies completed at this institution from 1991 to 1992. For all patients, needle localization and excisional biopsy were completed and histologic results were available. Of the 206 lesions evaluated there were 99 masses alone, 76 suspicious calcifications, and 11 combinations of masses and associated microcalcifications. The remaining 20 lesions included various combinations of architectural distortion, regions of asymmetric breast density, areas of focal asymmetric density, and areas of asymmetric breast tissue. Patients ranged in age from 24 to 86 years with an average age of 55 years. At biopsy, 133 (65%) of the lesions were found to be benign while 73 (35%) were malignant.

The mammographic findings were encoded using the Breast Imaging Reporting and Data System or BI-RADS, a standardized lexicon devised by the American College of Radiology to improve upon the consistency of mammographic reports [44,45]. Each set of films was reviewed prospectively by one of two radiologists whose primary clinical responsibilities are the interpretation of mammograms and the evaluation of breast lesions and who are familiar with the definitions of the BI-RADS descriptors. The radiologist was provided with the cranio-caudal and mediolateral-oblique views from both breasts, as well as any other

Table 1. Coding of findings to network input values

Calc. Distribution		Mass Margin		Location	
no calcifications	0	no mass	0	axillary tail	0
diffuse	0.2	well circumscribed	0.2	posterior	0.2
regional	0.4	microlobulated	0.4	middle	0.4
segmental	0.6	obscured	0.6	anterior	0.6
linear	0.8	ill-defined	0.8	subareolar	0.8
clustered	1.0	spiculated	1.0	central	1.0
Calc. Number		**Mass Size**		**Associated Findings**	
no calcifications	0	mm, rel. to max	%	none	0
< 5	0.33			skin lesion	0.11
5 to 10	0.66	**Mass Shape**		hematoma	0.22
> 10	1.0	no mass	0	post surgical scar	0.33
		round	0.25	trabecular thickening	0.44
Calc. Description		oval	0.5	nipple retraction	0.56
no calcifications	0	lobulated	0.75	skin retraction	0.67
milk of calcium-like	0.07	irregular	1.0	skin thickening	0.78
rim	0.14			architectural distortion	0.89
skin	0.21	**Mass Density**		axillary adenopathy	1.0
vascular	0.29	no mass	0		
spherical	0.36	fat-containing	0.25	**Special Cases**	
suture	0.43	low density	0.5	none	0
coarse	0.50	isodense	0.75	intramam. lymph node	0.25
large rod-like	0.57	high density	1.0	asym. breast tissue	0.5
round	0.64			focal asym. density	0.75
dystrophic	0.71			tubular density or	1.0
punctate	0.79			solitary dilated duct	
indistinct	0.86				
pleomorphic	0.93			**Age**	
fine branching	1.0			yrs, rel. to max	%

available views and films from prior studies. The radiologist was blinded to the biopsy result and reviewed the films prospectively.

The radiologist was asked to describe ten radiographic findings pertaining to lesion morphology. The first three are descriptive features that apply to microcalcifications and calcifications associated with masses: calcification distribution, number and description. Another four features apply only to masses: mass margin, mass shape, mass density, and mass size. Three features that can apply to all lesions include lesion location, associated findings (e.g. axillary adenopathy), and special cases (e.g. asymmetric breast tissue). The patient age was also recorded. The radiologist also assigned an overall impression of malignancy on a scale from one to five: one = benign, two = probably benign, three = indeterminate, four = probably malignant, and five = malignant. This estimate of probability for malignancy was used only to evaluate the radiologists' performance.

For quantitative features, the neural network input was simply the numeric value, such as mass size in millimeters normalized by the maximum mass size. In comparison, qualitative features were recorded in a multiple-choice format. Radiologists selected one of several possible descriptors for each feature, such as the six choices for mass margin: no mass, well circumscribed, microlobulated, obscured, indistinct, or spiculated. These feature descriptors were then coded into equally-spaced numeric values from zero to one, as shown in Table 1 above. The ordering of the descriptors for each feature was arrived at by discussion with experienced

mammographers and review of reports discussing the malignant potential of various BI-RADS descriptors [45,46].

2.2. Backpropagation neural network architecture

The ANN architecture we used were three-layer feed-forward backpropagation networks. A typical network is illustrated in Figure 2 on the next page. For this network, the input layer consists of eight input nodes (three features are excluded to simplify the figure). The inputs are fully connected to the hidden layer which in this case consists of 16 nodes. These hidden nodes are in turn connected to the output layer, consisting of a single node representing diagnostic outcome. The network's output is compared to the desired or target output, which is set to the actual pathologic diagnosis: 0 for benign and 1 for malignant.

For this features-to-diagnosis network, we use the subscript i for the input layer, j for the hidden layer, and k for the output layer. The weight connecting input feature F_i and hidden node H_j is W_{ij}, and the weight connecting hidden node H_j and the output node or diagnosis D_k is V_{jk}. Since the network has only a single output, the k subscripts are unnecessary and thus omitted.
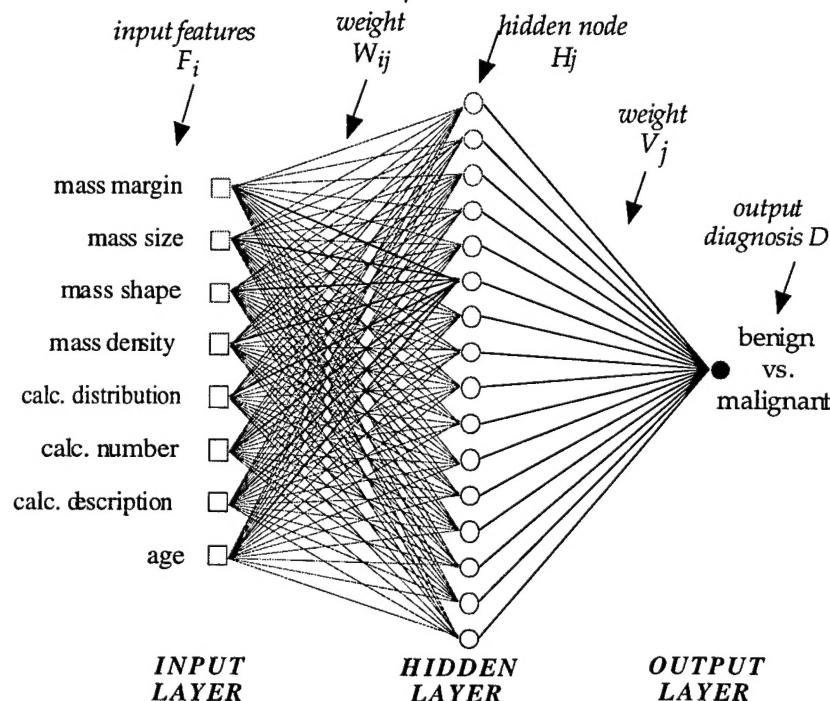

Figure 2. Preliminary neural network architecture.

Each node in the hidden and output layers calculate the weighted sum of its inputs from the previous layer, add a bias value, then pass the resulting sum through a sigmoid thresholding function (shown below) to yield an output value between 0 and 1.

$$f\{x\} = \frac{1}{1+e^{-x}} \qquad (1)$$

This process is illustrated in fig. 2 above by the "fan-in" of the darker lines (i.e., weights W_{i6}) to the 6th hidden node. The bias is added and the sum passed through the sigmoid function to produce the output value for that hidden node, H_j, as shown in Eq. 2 below.

$$H_j = f\left\{\sum_i^{N=8} W_{ij} F_i + bias_j\right\} \qquad (2) \qquad\qquad D = f\left\{\sum_j^{M=16} V_j H_j + bias_{out}\right\} \qquad (3)$$

The outputs from the sixteen hidden nodes then become inputs to the last layer, consisting of the single output node. As before, the weighted sum of these inputs is formed, added by the single bias value $bias_{out}$, and passed through the sigmoid function, as shown in Eq. 3 above. The resulting single output is the diagnosis D. The diagnosis lies between 0 and 1, and may be compared with a threshold (between 0 and 1) during evaluation. If the output exceeds the threshold, a malignancy is predicted, otherwise the outcome is predicted to be benign.

2.3. Backpropagation training algorithm

The "knowledge" of each ANN was contained in its weights, which were initialized to small random numbers (uniformly distributed between 0.3 and -0.3). The network "learned" by using the generalized delta rule, whereby it adapts those weights over many presentations of training cases or iterations, using a gradient descent technique to minimize the mean squared error between the network and target outputs.

Using the round robin or "leave one out" technique, the network was trained on all but one of the examples for a fixed number of iterations, then tested on the one excluded example [33]. The excluded example was replaced, the network weights were reinitialized, and the training was repeated by excluding a different example until every example had been excluded once.

For each training case consisting of the input/output pair p, the network output D was compared against the target diagnosis D_p, (set to one if the biopsy reveals malignancy and zero if benign) to form the error $Derr_p$:

$$Derr_p = D(1-D)(D_p - D) \qquad (4)$$

Using this error term, we calculated additive correction or "delta" factors for the output node's weights and bias at iteration n:

$$\Delta V_j^n = \eta \cdot H_j \cdot Derr_p + \alpha \cdot \Delta V_j^{n-1} \qquad\qquad \Delta bias_{out}^n = \eta \cdot Derr_p \qquad (5)$$

The delta factor for a weight depended on both the error and the input which caused that error. Both η and α were proportionality constants. The learning rate η controlled the rate of convergence, while the momentum α enhanced the speed and

stability of convergence by incorporating part of the old delta factor into the current one. The old weights and bias were added by their delta factors to yield the current weights and bias:

$$V_j^n = V_j^{n-1} + \Delta V_j^n \qquad\qquad bias_{out}^n = bias_{out}^{n-1} + \Delta bias_{out}^n \qquad\qquad (6)$$

Since there were no target values to compare against the hidden nodes' values, the output error $Derr_p$ was backpropagated to determine each hidden node's error $Herr_{jp}$:

$$Herr_{jp} = H_j (1 - H_j) V_j Derr_p \qquad\qquad (7)$$

As before, delta factors were calculated for the hidden nodes' weights and biases:

$$\Delta W_{ij}^n = \eta F_i Herr_{jp} + \alpha \Delta W_{ij}^{n-1} \qquad\qquad \Delta bias_j^n = \eta Herr_{jp} \qquad\qquad (8)$$

After each iteration, consisting of one complete presentation of all training cases, the training and testing MSE were calculated. The testing MSE over the L=260 cases was:

$$MSE^n = \frac{1}{L} \sum_p^L (D_p - D)^2 \qquad\qquad (9)$$

Also after each iteration, for the L testing outputs, the sensitivity and specificity over a range of decision thresholds were expressed as a receiver operating characteristic (ROC) curve. Performance of both the networks and radiologists were measured and compared by the ROC area index, Az, using LABROC4 and CLABROC software (provided by Dr. Charles Metz, University of Chicago, Chicago, IL) [47,48]. Large area indices close to 1 corresponded to high specificity and sensitivity.

The optimal number of iterations was found by halting training when the testing MSE no longer decreased, indicating when the network had become overtrained, thus losing its ability to generalize to new cases. Since the final measure of merit was Az, the effect of halting training when the Az no longer increased was also investigated. Minimizing the testing error almost always yielded the same stopping criterion as maximizing ROC area. The neural networks required 200–1000 iterations to minimize error. Maximizing Az instead sometimes increased or decreased training by a few hundred iterations, but never improved Az by more than 30% of a standard deviation.

Similarly, the number of hidden nodes were varied from 5 to 25 to optimize the Az. More hidden nodes would permit the network to form more complex decision regions and become a better classifier. Too many hidden nodes, however, would result in too many weights than can be reliably estimated from the limited number of examples [49]. The network was robust to variations in the number (between 5 and 25) of hidden nodes. Although it performed best with fifteen and worst with five hidden nodes, the difference in Az was only 1%. This trend was typical of all networks in this study. Therefore unless otherwise noted, all of the following results were reported for networks with fifteen hidden nodes.

All neural network software was custom-written by the principal investigator in the C language and executed on SPARC 20 workstations (Sun Microsystems, Inc., Mountain View, CA). For each run consisting of a new combination of inputs and hidden nodes, the round robin process of training and testing required approximately 4 hours without parallelization.

2.4. Optimized reduction of input features

Given all ten mammographic features and patient age, the best 11-feature network performed with Az of 0.84 ± 0.03, which was not significantly different from the expert radiologists' Az of 0.85 ± 0.03 (2-tailed p-value = 0.54). This was an important result since the network did not have access to all the information that radiologists did, such as mammograms from other views and previous studies, clinical findings, and history findings other than age. In a separate paper, we demonstrated further improvements by including history findings in the ANN inputs, raising Az to 0.89 ± 0.02 which was still not significantly better than the radiologists (p=0.29) [39].

We next sought to identify the minimal subset of features which would still yield accurate diagnostic performance. There were several motivating reasons for doing so. Fewer features would reduce the data-entry effort of radiologists, which in turn makes it more likely that they would incorporate the ANN into their standard reading process. Previous studies have shown that community radiologists and technologists may extract features as reliably as expert mammographers, but lack the latter's experience in merging those features into a diagnosis [50,51]. Simplifying the number of inputs may enable the use of community radiologists or technologists for feature extraction, thus improving accessibility of expert diagnosis. Fewer inputs should also permit reducing the number of hidden nodes and hence the number of ANN weights, thus ameliorating the problem of overconditioning due to insufficient training cases [49]. Finally, a simplified computer model may shed some light on the complex cognitive processes underlying radiological diagnosis.

In general, input features may be eliminated one or a few at a time, then the algorithm retrained and retested to determine the significance of the excluded feature(s). Since many features are correlated, however, this process becomes more complicated. In other words, groups of features may be greater or less than the sum of their parts. Previous authors reduced the number of inputs by using only those inputs whose mean value differed greatly for benign vs. malignant cases [33]. Others employed statistical techniques such as linear discriminant analysis to identify an optimal subset of inputs [52]. To fulfill the objectives of this study, we employed a new technique using nonlinear ANNs to identify the optimized subset of features.

Since each subset of features required developing a new ANN (varying the number of hidden nodes and then performing a round robin for each), it was impractical to investigate all possible combinations. Instead, the features were ordered by their importance, then eliminated one by one until network performance was significantly degraded.

To rank the features by importance, one feature was excluded and a neural network was retrained using all the other features. As before, network performance was measured by Az. The process was then repeated with a different feature until all features had been excluded once. The assumption was that the exclusion of an important feature would degrade performance more than the exclusion of an unimportant feature. In order from most to least important, the features were: (1) age, (2) mass margin, (3) calcification description, (4) mass density, (5) associated findings, (6) calcification distribution, (7) lesion location, (8) mass size, (9) calcification number, (10) mass shape, and (11) special cases. Excluding the most important feature (age) reduced Az from 0.84 to 0.80, while excluding the least important feature (special cases) did not affect Az at all.

Figure 3. Reducing number of features
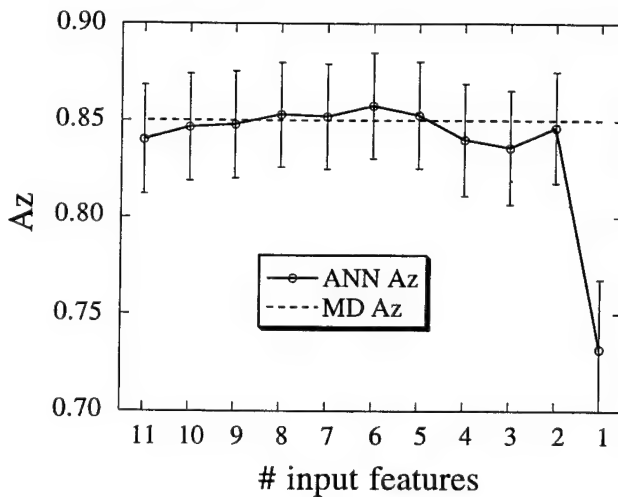
As ANN features are reduced from 11 to 2, Az is comparable to radiologists (dashed line).

Table 2. Performance of ANNs as the number of features are reduced

# features	Az	σ	p
11	0.84	0.03	0.54
6	0.86	0.03	0.43
2	0.85	0.03	0.83
radiologists	0.85	0.03	–

For each network, the Az and standard deviations are shown, along with the p-value for the difference compared to radiologists. The 6-feature network is the best, but all ANNs showed no difference vs. radiologists.

Once ranked, the input features were discarded in order from least to most important in a manner analogous to backwards discrimination analysis, reducing the number of features to ten, nine, eight, and so on. Each simplified network was retrained and re-tested with the round robin process as before, and its performance was compared to that of the expert radiologists. The performance of these simplified networks are plotted in Figure 3 and summarized in Table 2.

As the number of features was reduced from eleven to only two, the Az barely fluctuated with changes of much less than one standard deviation. Even in the extreme case, the two-feature network still performed well with Az of 0.85 ± 0.03, which was not statistically significantly different compared to the radiologists (p = 0.83). The six-feature network emerged as the best compromise between minimizing features and maximizing performance. Its ROC area of 0.86 ± 0.03 was not

Fig. 4. ROC curves of 6-feature ANN vs. radiologists.

significantly different than that of the expert radiologists with p = 0.34. The ROC curve of the six-feature network is plotted in Figure 4 against that of the radiologists.
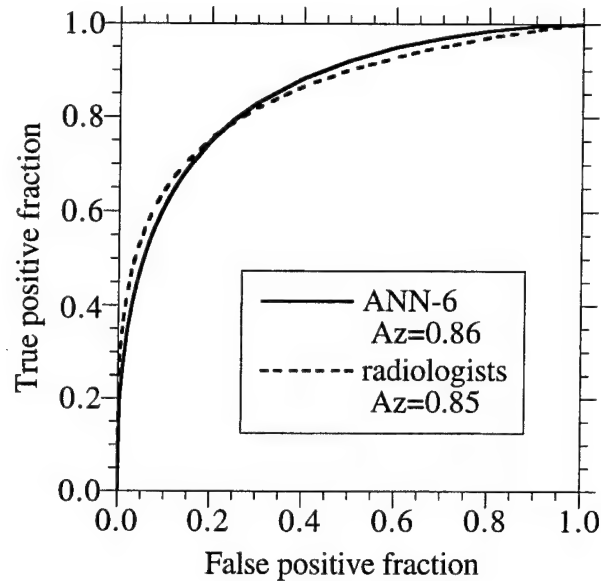
Histograms of the neural network outputs and radiologist impressions for all cases are plotted as Figures 5 and 6, respectively. Note the gaussian shape of the radiologist impressions where most cases were indeterminate, compared to the bimodal outputs of the computer models where most cases could be definitively diagnosed as positive or negative.
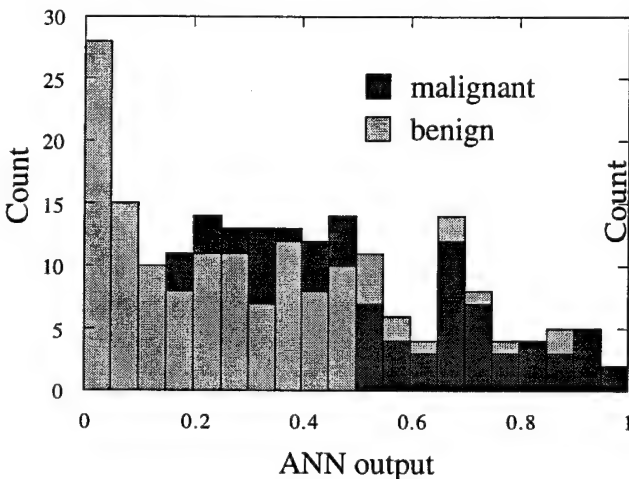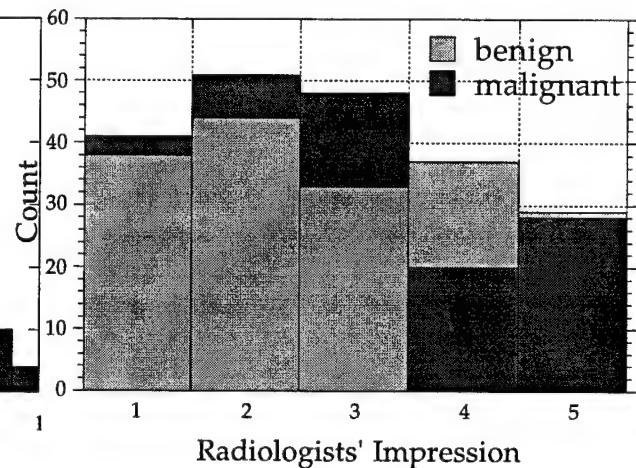
Fig. 5. Histogram of ANN outputs

Fig. 6. Histogram of radiologists impressions

In addition to high ROC area, it is also important to have good specificity at the very high sensitivities required in clinical practice. We considered the particular operating point on each ROC curve where sensitivity was 95%, corresponding to missing 5% of cancers of the malignancies identified by radiologists. At 95% sensitivity, the specificity of the neural network (56%) was statistically significantly greater than the specificity of the radiologists alone (30%) with two-tail p-value < 0.01. In terms of actual patients in this study, at 95% sensitivity the ANN would have missed 3 out of 73 malignancies but prevented 74 of 133 benign biopsies. At the portion of the ROC curves where most mammographers are trained to practice (i.e. high sensitivity and low specificity) the ANN maintained a very high relative sensitivity while significantly improving the mammographer's specificity.

This work demonstrated an empirical technique for identifying an optimal subset of input features to a complex, nonlinear classification system. There was minimal change in network performance as the number of inputs was pared from eleven down to six, which represented an optimal compromise between minimizing the number of input features and maximizing performance. The ANN's optimized subset of features correlated well with those identified by expert radiologists as being among the most important, although no clinician would be willing to make a diagnosis based on so few findings. It was therefore all the more remarkable that the neural network outperformed the specificity of the expert mammographers who extracted the input features in the first place and who also had access to other information such as previous films and the patients' clinical history.

2.5 Other technical objectives

The preceding sections pertained to specific aim 1a listed in the proposal's technical objectives. In specific aim 1b, we originally proposed to separate the features identified in aim 1a into binary sub-features, thus facilitating their automated extraction. Preliminary work during this first year suggested several reasons not to focus so specifically on individual sub-features. First, analysis of the distribution of the various sub-features revealed that some sub-features were present in very few patients, which would make teach-by-example development of ANNs difficult. Second, the original proposal assumed that each feature was separable into discrete, non-overlapping sub-features. Our studies demonstrated considerable inter-observer variation in categorizing each feature into sub-features [40]. Finally, for some features, the sub-features may be considered as approximate gradations of the same phenomenon. For example, the mass margin ranges from circumscribed which is very "smooth" to spiculated which is very "rough." For these reasons, we deemed it unnecessary to separate these inter-related, overlapping sub-features for ANN development.

Work on specific aim 2a was commenced in accordance with the time line shown in the original proposal. The initial results will be presented at a national conference during the second budget period [36], so discussion of that work will be reserved for the second annual report.

3.   Conclusions

This goal of this proposal is to develop a computer-aided diagnosis system to automatically extract radiographic features from the mammogram, then use an artificial neural network (ANN) to merge those features to predict breast lesion malignancy. During the first budget period, we successfully developed an ANN that merges radiologist-extracted features to predict malignancy. By adopting the nationally-standardized BI-RADS lexicon for encoding features, this ANN has the potential to be widely applicable. Finally, we also developed an empirical technique for identifying an optimal subset of input features to a complex, nonlinear classification system. This technique can be applied to any problem where one wishes to optimally simplify a large number of input features for ANN development.

At the conclusion of the first budget period, we have accomplished the stated goals of the proposal for that time period, or indicated our reasons for not doing so. In the next year, we will continue in accordance with the proposal time line. Specifically, we will conclude work on aim 2a concerning conventional techniques for automated feature extraction. We will at the same time commence work on aim 2b, using ANNs for feature extraction. If successful, the features extracted by conventional or ANN methods will eventually be fed to the features-to-diagnosis network developed during the first year. Together, the system will provide automated, accurate predictions of breast lesion malignancy.

4. References

1   Boring CC, Squires TS, Tong T, and Montgomerie S, "Cancer Statistics 1994," *CA Cancer J. Clin.* 44, 7-26 (1994).

2   Shapiro S, "Screening: assessment of current studies," *Cancer* 74, 231-238 (1994).

3   Rutqvist LE, Miller AB, Andersson I, Hakama M, Hakulinen T, Sigfusson BF, and Tabar L, "Chapter IV: Reduced breast cancer mortality with mammography screening - an assessment of currently available data," *International Journal of Cancer* Supplement 5, 76-84 (1990).

4   Kopans DB, "The positive predictive value of mammography," *AJR* 158, 521-526 (1992).

5   Knutzen AM, and Gisvold JJ, "Likelihood of malignant disease for various categories of mammographically detected, nonpalpable breast lesions," *Mayo Clinic Proceedings* 68, 454-460 (1993).

6   Bassett LW, Bunnell DH, Cerny JA, and Gold RH, "Screening mammography: referral practices of Los Angeles physicians," *Am J Roentgenol* 147, 689-92 (1986).

7   Cyrlak D, "Induced costs of low-cost screening mammography," *Radiology* 168, 661-3 (1988).

8   Varas X, Leborgne F, and Leborgne JH, "Nonpalpable, probably benign lesions: role of follow-up mammography," *Radiology* 184, 409-14 (1992).

9   Sickles EA, "Periodic mammographic follow-up of probably benign lesions: results in 3,184 consecutive cases," *Radiology* 179, 463-8 (1991).

10 Doi K, Giger ML, MacMahon H, Nishikawa RM, Schmidt RA, and Hoffman KR, "Computer-aided diagnosis: Potential usefulness of real-time computer output to radiologists' interpretations," *Radiology* 185(P), 391 (1992)(abstract).

11 Doi K, Giger ML, MacMahon H, Hoffmann KR, Nishikawa RM, Schmidt RA, Chua K-G, Katsuragawa S, Nakamori N, Sanada S, Yoshimura H, Metz CE, Montner SM, Matsumoto T, Chen X, and Vyborny CJ, "Computer-aided diagnosis: Development of automated schemes for quantitative analysis of radiographic images," *Sem. Ultrasound CT MRI* 13, 140-152 (1992).

12 Asada N, Doi K, MacMahon H, Montner SM, Giger ML, Abe C, and Wu Y, "Potential usefulness of an artificial neural network for differential diagnosis of interstitial lung diseases: Pilot study," *Radiology* 177, 857-860 (1990).

13 Katsuragawa S, Doi K, Nakamori N, and MacMahon H, "Image feature analysis and computer-aided diagnosis in digital radiography: Effect of digital parameters on the accuracy of computerized analysis of interstitial disease in digital chest radiographs," *Med. Phys.* 17, 72-78 (1990).

14 Nakamori N, Doi K, Sabeti V, and MacMahon H, "Image feature analysis and computer-aided diagnosis in digital radiography: Automated analysis of sizes of heart and lung in digital chest images," *Med. Phys.* 17, 342-350 (1990).

15 Sanada S, Doi K, and MacMahon H, "Image feature analysis and computer-aided diagnosis in digital radiography: Automated detection of pneumothorax in chest images," *Med. Phys.* 19, 1153-1160 (1992).

16 Giger ML, Ahn N, Doi K, MacMahon H, and Metz CE, "Computerized detection of pulmonary nodules in digital chest images: Use of morphological filters in reducing false-positive detections," *Med. Phys.* 17, 861-865 (1990).

17 Lo SB, Freedman MT, Lin JS, and Mun SK, "Automatic lung nodule detection using profile matching and back-propagation neural network techniques," *J. Digit. Imaging* 6, 48-54 (1993).

18 Garg S, and Floyd CE, Jr, "Neural network localization of pulmonary nodules on digital chest radiographs," *Radiology* 185(P), 157 (1992)(abstract).

19 Tourassi GD, and Floyd CE, Jr, "Artificial neural networks for Single Photon Emission Computed Tomography: A study of cold lesion detection and localization," *Invest. Radiol.* 28, 671-677 (1993).

20 Tourassi GD, Floyd CE, Jr, Sostman HD, and Coleman RE, "Acute pulmonary embolism: Artificial neural network approach for diagnosis," *Radiology* 189, 555-558 (1993).

21 Chan HP, Doi K, Vyborny CJ, Lam KL, and Schmidt RA, "Computer-aided detection of microcalcifications in mammograms: Methodology and preliminary clinical study," *Invest. Radiol.* 23, 664-671 (1988).

22 Ng SL, and Bischof WF, "Automated detection and classification of breast tumors," *Computers and Biomedical Research* 25, 218-237 (1992).

23 Yin FF, Giger ML, Vyborny CJ, Doi K, and Schmidt RA, "Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses," *Invest. Radiol.* 28, 473-481 (1993).

24 Lucier BJ, Kallergi M, Qian W, DeVore RA, Clark RA, Saff EB, and Clarke LP, "Wavelet compression and segmentation of digital mammograms," *J. Digit. Imaging* 7, 27-38 (1994).

25 Qian W, Clarke LP, Li HD, Clark R, and Silbiger ML, "Digital mammography: M-channel quadrature mirror filters (QMFs) for microcalcification extraction," *Comput Med Imaging Graph* 18, 301-14 (1994).

26 Wu Y, Doi K, Giger ML, and Nishikawa RM, "Computerized detection of clustered microcalcifications in digital mammograms: Applications of artificial neural networks," *Med. Phys.* 19, 555-560 (1992).

27 Zhang W, Doi K, Giger ML, Wu Y, Nishikawa RM, and Schmidt RA, "Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network," *Med. Phys.* 21, 517-524 (1994).

28 Parker J, Dance DR, Davies DH, Yeoman LJ, Michell MJ, and Humphreys S, "Classification of ductal carcinoma in situ by image analysis of calcifications from digital mammograms," 68, 150-9 (1995).

29 Walls J, Boggis CR, Wilson M, Asbury DL, Roberts JV, Bundred NJ, and Mansel RE, "Treatment of the axilla in patients with screen-detected breast cancer," *Br J Surg* 80, 436-8 (1993).

30 Freundlich IM, Hunter TB, Seeley GW, D'Orsi CJ, and Sadowsky NL, "Computer-assisted analysis of mammographic clustered calcifications," *Clin. Radiol.* 40, 295-298 (1989).

31 Patrick EA, Moskowitz M, Mansukhani VT, and Gruenstein EI, "Expert learning system network for diagnosis of breast calcifications," *Invest. Radiol.* 26, 534-539 (1991).

32 Giger ML, Vyborny CJ, and Schmidt RA, "Computerized characterization of mammographic masses: analysis of spiculation," *Cancer Letters* 77, 201-211 (1994).

33 Wu Y, Giger ML, Doi K, Vyborny CJ, Schmidt RA, and Metz CE, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," *Radiology* 187, 81-87 (1993).

34 Rosenblatt F, "The Perceptron: A probabilistic model for information storage and organization in the brain," *Psych. Rev.* 65, 386-408 (1958).

35 White H, "Learning in artificial neural networks: a statistical approach," *Neural Comput* 1, 425-464 (1989).

36 Lo JY, Baydush AH, Baker JA, Kornguth PJ, and Floyd CE, Jr, "Computer-aided diagnosis of breast mass malignancy using automated feature extraction and artificial neural networks," *Radiology* 197(P), (1995)(abstract).

37 Floyd CE, Jr, Lo JY, Yun AJ, Sullivan DC, and Kornguth PJ, "Prediction of breast cancer malignancy using an artificial neural network," *Cancer* 74, 2944-2948 (1994).

38 Lo JY, Floyd CE, Jr, and Tourassi GD, "Artificial neural networks for diagnosis in radiology," in Computer-Aided Diagnosis Workshop, Georgetown University Medical Center, Washington, DC (Georgetown University Medical Center, Washington, DC, 1994), p.22.

39 Baker JA, Kornguth PJ, Lo JY, Williford ME, and Floyd CE, Jr, "Artificial neural network for the prediction of breast cancer using BI-RADS™ standardized lexicon," *Radiology* 196, 817-822 (1995).

40 Baker JA, Kornguth PJ, Lo JY, and Floyd CE, "An artificial neural network approach to improving the quality of breast biopsy recommendations," *Radiology* accepted, (1995).

41 Lo JY, Baker JA, Kornguth PJ, and Floyd CE, Jr, "Computer-aided diagnosis of mammography: Artificial neural networks for optimized merging of standardized BIRADS features," in World Congress on Neural Networks 95 (International Neural Network Society Annual Meeting), Washington, D.C. (Washington, D.C., 1995), p.885-888.

42   Lo JY, Grisson AT, Floyd CE, Jr, and Kornguth PJ, "Computer-aided diagnosis of mammograms using an artificial neural network: merging of standardized input features from the ACR lexicon," in SPIE Medical Imaging 1995: Image Processing, (1995), p.571-578.

43   Lo JY, Baker JA, Kornguth PJ, and Floyd CE, Jr, "Computer-aided diagnosis of breast cancer: artificial neural network approach for optimized merging of mammographic features," *Acad. Radiol.* accepted, (1995).

44   Kopans DB, "Standardized mammography reporting," *Radiologic Clinics of North America* 30, 257-264 (1992).

45   D'Orsi CJ, and Kopans DB, "Mammographic feature analysis," *Seminars in Roentgenology* 28, 204-230 (1993).

46   BIRADS. Reston, VA: American College of Radiology. *Breast Imaging - Reporting and Data System (BI-RADS).* 1993.

47   Metz CE, "Some practical issues of experimental design and data analysis in radiological ROC studies," *Invest. Radiol.* 24, 234-245 (1989).

48   Rockette HE, Gur D, and Metz CE, "The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques," *Invest. Radiol.* 27, 169-172 (1992).

49   Lippman RP, "An introduction to computing with neural nets," *IEEE ASSP Magazine* 4-22 (1987)(review article).

50   Getty DJ, Pickett RM, D'Orsi CJ, and Swets JA, "Enhanced interpretation of diagnostic images," *Invest. Radiol.* 23, 240-252 (1988).

51   Bassett L, Hollatz-Brown AJ, Bastani R, Pearce JG, Hirji K, and Chen L, "Effects of a program to train radiologic technologists to identify abnormalities on mammograms," *Radiology* 194, 189-192 (1995).

52   Gale AG, Roebuck EJ, Riley P, and Worthington BS, "Computer aids to mammographic diagnosis," *British J. Radiol.* 60, 887-891 (1987).